Thomas **Winninger**

Student at Télécom SudParis on a gap year

thomas.winninger@telecom-sudparis.eu https://le-magicien-quantique.github.io Sckathach Thomas Winninger
0009-0000-2783-3086

Whoami?

Aka *the quantum warlock, the masked camel, the fanOfThermodynamics, the pipe and clouds engineer, the whale orchestra conductor*, or just **Sckathach**. I'm a french student at Télécom SudParis, and soon, a ML(4SEC) researcher!
Fond of mathematics and physics, I ended up at the Télécom SudParis engineering school where I focused on cyber security. As I quickly became interested in AI security, I decided to take a gap year to bring myself up to speed on the subject: AI security research, interpretability, tools, statistics; and since that's what I like best, I plan to continue with a master's degree and a thesis, most likely in the same field.

Education

Master's Degree in Computer Science - Cyber Security?

Télécom SudParis - Institut polytechnique de Paris (IPP) 2024 - 2026

Engineering Degree - Cyber Specialization

Télécom SudParis 2022 - 2026

Telecommunications, network security and web applications, graph theory (application to AI and 6G). Computer science theory and databases. Signal processing and probability.

Experience

Research Internship in Language Model Explainability

INRIA - ANTIQUE March - May 2025

Language model explainability through abstract interpretation.

Research Internship in AI Security

Thales - ThereSIS July - December 2024

Implementation and improvement of state-of-the-art attacks on LLMs.

Training/Infrastructure Manager

HackademINT 2023 - 2024

Creation of challenges (AI & quantum physics), and organization of 404CTF 2023 & 2024.

Talks

- **Mechanistic interpretability for LLM attack and defense** - *École Polytechnique, CeSIA (avril 2025)*
- **Introduction to AI security and reverse engineering** - *HackademINT (avril 2025)*
- **Model Poisoning** - *AI Safety Meetup | Centre pour la sécurité de l'IA (CeSIA) (juin 2024)*
- **Détection de la triche dans le 404 CTF** - *Rendez-vous de la Recherche et de l'Enseignement de la Sécurité des Systèmes d'Information* (mai 2024)

Papers

- **Using Mechanistic Interpretability to craft Adversarial Attacks against Large Language Models** - *Winninger T., Addad B., Kapusta K.* (mars 2025)

Skills

Programming Languages	**Python**, **Ocaml**, TypeScript, Typst, Rust, Lua, C, Bash
Spoken Languages	**French**, **English**, Korean, Japanese
Tools	**PyTorch**, PyG, Docker (Podman), Kubernetes, React, Qiskit, Sage, Archlinux :)

Other Interests

Piano, guitar, video game development, reading, geopolitics, particle physics :), sports, meditation, teaching.